

AD-A079 418

CENTER FOR NAVAL ANALYSES ALEXANDRIA VA  
NONPARAMETRIC METHODS FOR ESTIMATING RECRUIT SURVIVAL WITH CROS--ETC(U)  
SEP 79 P M LURIE  
N00014-76-C-0001

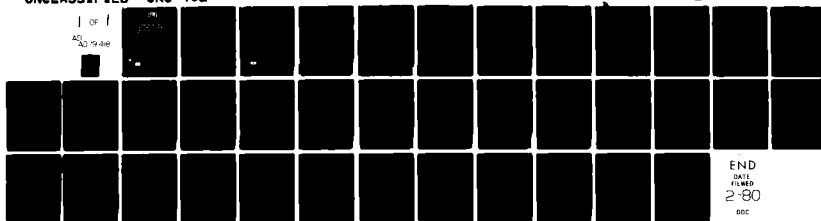
F/9 12/1

UNCLASSIFIED

CRC-402

NL

1 OF 1  
AD-A079 418



CRC 402 / September 1979

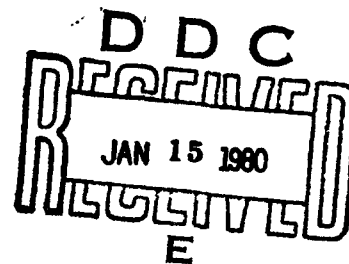
**LEVEL**

12  
F

ADA 079418

# NONPARAMETRIC METHODS FOR ESTIMATING RECRUIT SURVIVAL WITH CROSS-SECTIONAL DATA

Philip M. Lurie



DDC FILE COPY

Approved for public release;  
distribution unlimited.



**CENTER FOR NAVAL ANALYSES**

2000 North Beauregard Street, Alexandria, Virginia 22311

80 1 14 017

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER CRC-442	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Nonparametric Methods for Estimating Recruit Survival With Cross-sectional Data		5. TYPE OF REPORT & PERIOD COVERED
7. AUTHOR(s) Philip M. Lurie		6. PERFORMING ORG. REPORT NUMBER
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Naval Analyses 2000 N. Beauregard Street Alexandria, Virginia 22311		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0001
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Department of the Navy Arlington, Virginia 22217		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 12 38
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Office of the Chief of Naval Operations (Op96) Department of the Navy Washington, D.C. 20350		12. REPORT DATE Sept 1979
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		13. NUMBER OF PAGES 34
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		15. SECURITY CLASS. (of this report) Unclassified
18. SUPPLEMENTARY NOTES This Research Contribution does not necessarily represent the opinion of the Department of the Navy.		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) estimates, logit analysis, models, probit analysis, recruiting, regression, statistical analysis, survival (general)		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) A survey of nonparametric methods (methods which make no distributional assumptions about the data) for survival curve estimation is presented. This is provided as background to the discussion of the Cox regression model, which can be applied to cross-sectional data. The Cox model is then compared to probit analysis on the 1973 recruit cohort of four-year obligors. Evidence is presented to show that the Cox model can be useful for estimating recruit survival from cross-sectional data.		

125

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE  
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

077 270

mt

CRC 402 / September 1979

# NONPARAMETRIC METHODS FOR ESTIMATING RECRUIT SURVIVAL WITH CROSS-SECTIONAL DATA

Philip M. Lurie



*Institute of Naval Studies*

**CENTER FOR NAVAL ANALYSES**

2000 North Beauregard Street, Alexandria, Virginia 22311

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DDC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/ _____	
Availability Codes	
Dist	Avail and/or special
<i>A</i>	

80

1 14 017

#### ACKNOWLEDGMENTS

The author would like to thank M. Pagano and I. Chang of the Sidney Farber Cancer Institute for providing the computer program on which the Cox regression analysis in this report is based.

## TABLE OF CONTENTS

	Page
Summary.....	iii
Introduction.....	1
Longitudinal versus cross-sectional data.....	3
Nonparametric estimation of a survival curve with-	
out covariates.....	5
Empirical distribution function.....	5
Product-Limit estimator.....	6
Life Table analysis.....	9
Nonparametric estimation of a survival curve with	
covariates.....	16
Probit and logit analyses.....	16
The Cox regression model.....	17
Comparison of the Cox and probit models.....	22
References.....	32

## SUMMARY

In this paper, we present a survey of some of the most commonly used nonparametric<sup>1</sup> methods of survival curve estimation. This is provided as background to the discussion of a relatively recent technique which can be applied to cross-sectional data. The method, known as Cox regression, has been used extensively in the biomedical sciences for relating covariates to patient survival, but to our knowledge has never been applied to military manpower problems. A prime objective of this paper is to examine the potential usefulness of this procedure by applying it to the 1973 recruit cohort of four-year obligors. This data set has been analyzed previously by means of a probit analysis, and the results are used as a basis for comparison with the Cox procedure.

A favorable comparison of the Cox procedure with probit analysis on a longitudinal data base (the 1973 cohort) will be regarded as evidence of its potential application to cross-sectional data, to which probit analysis cannot be applied. Furthermore, the results of an analysis using the Cox model are more easily interpretable and computationally more efficient than the probit counterparts. On the evidence presented in this paper, the Cox model appears to be a very useful procedure for estimating recruit survival from cross-sectional data.

---

<sup>1</sup>The term "nonparametric" means that no assumptions are made about the mathematical form of the distribution of the data.

## INTRODUCTION

Extensive analyses have been done by CNA which relate various pre-service and in-service personnel characteristics to the probability of surviving to a given point in time (usually the end of the first year or enlistment term). Statistical techniques which have been employed for identifying these relationships include probit and logit analysis (see references 1 and 2 for discussions of these methods). These techniques require a sample or population of individuals followed from the day they enter the Navy until they either leave or complete their first term (longitudinal data). Thus if we are considering 4-year obligors, for instance, it will be necessary to follow a cohort through 4 years of service.

In order to avoid following individuals for such a long period of time, it has been proposed that cross-sectional data be used for estimation of survival. A cross-sectional data set is formed by selecting all those individuals who are currently enlisted in the Navy. This provides an enormous and potentially very informative data base. However, the statistical methods mentioned above cannot be used to exploit this type of data, and therefore a different method must be employed.

The statistical technique which we propose to use for handling cross-sectional data is termed a Cox regression model (see reference 3). This model is used extensively in the biological and health sciences for relating covariates to survival of patients, but to our knowledge has never been applied to military manpower problems. The Cox model has the advantage of being able to generate a continuous survival curve rather than just a point estimate (which a probit analysis gives). It should be noted that an approximation to a continuous survival curve can be obtained by applying a probit analysis in a sequential manner, e.g., at monthly intervals, but only with a great deal of computational time and expense. The Cox model, on the other hand, can generate a survival curve at only a fraction of the time and cost.

The Cox model has the additional advantage that it can be applied to cross-sectional data, but it can, of

course, also be applied to longitudinal data. Thus we can compare the relative estimating abilities of probit and Cox on data from one cohort and, from the results, make recommendations as to the efficacy of using the Cox procedure in the future. We shall base the comparisons on the 1973 recruit cohort of 4-year obligors. Separate analyses will be performed for GENDETs and A-schoolers classified according to mental group and education with their age, race, and primary dependents held constant.

## LONGITUDINAL VERSUS CROSS-SECTIONAL DATA

The advantages and disadvantages of using longitudinal and cross-sectional data are summarized below.

- Longitudinal data

- Advantages:

All individuals are observed until completion of term or attrition.

All individuals start in the same year and are subject to the same patterns of attrition.

- Disadvantages:

Data must be collected and followed for 4 years.

Analyses based on these data may no longer be current, i.e., present attrition patterns may be different from those observed 4 years ago.

- Cross-sectional data

- Advantages:

Attrition patterns observed are the most current.

The data need only be followed for a relatively short period of time (say 1 year).

- Disadvantages:

Not all individuals are observed until completion of term or attrition.

The data consist of a mixture of different cohorts.

Only the portion of a cohort still in the Navy at the time of data collection is observed.

The last disadvantage listed for cross-sectional data can present considerable problems of bias (in estimating survival probabilities) if one is not careful with the analysis. For example, if the data consist of  $n_1$  individuals with 1 year of service and  $n_2$  individuals with two years of service at the time of data collection and the data were followed for 1 year, then a greatly inflated estimate of the probability of surviving 1 year could be obtained: each of the  $n_2$  individuals survived one year, but the remainder of their cohort who left before this time has not been included in the sample. This difficulty can be overcome by performing a conditional analysis. If we let  $P(T \geq i)$  be the probability that an individual survives at least  $i$  years,  $i=1,2$ , then we may write

$$P(T \geq 2) = P(T \geq 2 \mid T \geq 1) \times P(T \geq 1). \quad (1)$$

The probability of surviving 1 year may then be estimated by considering only those who start their service at the time of data collection and observing their status at the end of 1 year. We can also estimate the conditional probability of surviving 2 years given that 1 year has already been completed, by considering only the  $n_1$  individuals and observing their status after 1 year of follow-up. The probability of surviving 2 years may then be obtained from expression (1). The same logic may of course be extended to more than 2 years. Thus the potential biases due to using cross-sectional data can be eliminated by a careful selection of subsets of the population at each point of time to be estimated.

Another feature of cross-sectional data is that after the period of follow-up there will be individuals who are still in their first term of enlistment, i.e., they have neither completed their first term nor left. Data such as these are termed in the biostatistical literature censored observations. Thus, if we are to use cross-sectional data for making inferences, we must use statistical methods which take censored data into account.

## NONPARAMETRIC ESTIMATION OF A SURVIVAL CURVE WITHOUT COVARIATES

Methods for handling censored observations for the estimation of survival curves are well documented, both parametric and nonparametric. We shall concern ourselves here, however, with the discussion of only the nonparametric methods.

A summary of the most commonly used nonparametric methods for estimating survival probabilities is presented below. Also given are the types of data to which the methods apply.

- Empirical Distribution Function (EDF) - no censored observations.
- Kaplan-Meier or Product-Limit (P-L) - generalization of EDF to censored data.
- Life Table - grouped data, interval counts; can handle censored data.

This is not an all-inclusive list, but other methods for estimating survival curves usually involve only modifications of those listed above.

### EMPIRICAL DISTRIBUTION FUNCTION

The Empirical Distribution Function (EDF) gives the nonparametric maximum likelihood estimate of the true underlying continuous distribution function  $F$ . If we observe data  $X_1, X_2, \dots, X_n$  which are independently and identically distributed (i.i.d.), then the EDF is defined as follows:

$$F_n(t) = (\# \text{ of } X_j \text{'s } \leq t) / n \quad (2)$$

$$= \frac{1}{n} \sum_{j=1}^n I(X_j - t) \quad , \quad (3)$$

where

$$I(X_j - t) = \begin{cases} 1 & \text{if } X_j - t \leq 0 \\ 0 & \text{if } X_j - t > 0 \end{cases} .$$

From expression (3) we see that the EDF is just a step function with jumps of  $1/n$  at each of the  $n$  order statistics (i.e., the ordered values of  $X_1, \dots, X_n$ ). Also from expression (3) we have that

$$\begin{aligned} E[F_n(t)] &= \frac{1}{n} \sum_{j=1}^n E[I(X_j - t)] \\ &= \frac{1}{n} \sum_{j=1}^n P(X_j \leq t) \\ &= \frac{1}{n} \sum_{j=1}^n F(t) \quad (\text{since } X_1, \dots, X_n \text{ are i.i.d.}) \\ &= \frac{1}{n} \cdot n F(t) \\ &= F(t), \end{aligned}$$

so that the EDF is seen to be an unbiased estimate of the true distribution function  $F$ . From the Strong Law of Large Numbers (see reference 4, for example), we also have that  $F_n(t)$  is a consistent estimator for  $F(t)$ ; i.e., for each value of  $t$ ,

$$\lim_{n \rightarrow \infty} P\{F_n(t) = F(t)\} = 1.$$

Thus as the number of observations becomes large,  $F_n(t)$  approximates a smooth function and approaches the true distribution function  $F(t)$ .

#### PRODUCT-LIMIT ESTIMATOR

The Product-Limit (P-L) estimator generalizes the EDF to handle censored observations. It was derived by Kaplan and Meier (reference 5) and was shown to be the non-parametric maximum likelihood estimate of the true survival function in the presence of censoring. To define

the P-L estimator, we consider the situation where failure times are observed at  $k$  distinct points  $0 < t_1 < t_2 < \dots < t_k$  with  $n_j$  failures occurring at each  $t_j$ . This allows for "discretizing" a continuous distribution; for example when service time is measured in months there will be many individuals with the same length of service. Now we also observe  $m_j$  censored observations in the interval  $I_j = [t_{j-1}, t_j)$ ,  $j=1, 2, \dots, k+1$ , where  $t_0 = 0$ ,  $t_{k+1} = \infty$ .

Kaplan and Meier showed that the maximum likelihood estimator of the survival function puts probability mass not at any censored observations, but only at the observed failure times. The censored observations do play a role, however, in determining what the probabilities at the failure times will be.

Now let  $r_j = \sum_{i=j}^k (n_i + m_{i+1})$ ,  $j = 1, \dots, k$ . This may be seen to be the number of observations that occur at times greater than or equal to  $t_j$ . Then if we let  $p_j = P(T > t_j \mid T > t_{j-1})$ , the maximum likelihood estimate of  $p_j$  is

$$\hat{p}_j = \frac{r_j - n_j}{r_j}, \quad j = 1, \dots, k. \quad (4)$$

The estimate  $\hat{p}_j$  is just the number of observations strictly greater than  $t_j$  divided by the number greater than or equal to  $t_j$ . Note that this estimate of  $p_j$  utilizes the censored observations beyond time  $t_j$  since they have definitely survived the interval  $I_j = [t_{j-1}, t_j)$ , but it essentially ignores those censored observations which occur within  $I_j$ . By virtue of ignoring them, we are, in effect, assigning  $\hat{p}_j$  proportion of them as survivors of the interval.

From (4) we may define the P-L estimator of  $S(t) = P(T > t)$  as

$$\hat{S}(t) = \begin{cases} 1 & \text{if } 0 \leq t < t_1 \\ \prod_{i=1}^j \hat{p}_i & \text{if } t_j \leq t < t_{j+1}, j=1, \dots, k-1 \\ \prod_{i=1}^k \hat{p}_i & \text{if } t_k \leq t < s^* \\ \text{undefined} & \text{for } t \geq s^* \end{cases} \quad (5)$$

where  $s^*$  represents the last observed data point if it is censored. Since the P-L estimator puts all its probability mass at the failure times, no additional mass is placed in the interval  $t_k \leq t < s^*$  other than at  $t_k$ . Although  $\hat{S}(t)$  is undefined for  $t \geq s^*$ ,  $\prod_{i=1}^k \hat{p}_i$  is, of course, an upper bound for  $\hat{S}(t)$  for  $t_k \leq t < \infty$ .

For an illustration of the P-L estimator, consider the following hypothetical data:

failure times at 0.8, 3.1, 5.4, 9.2  
censored times at 1.0, 2.7, 7.0, 12.1.

For this situation, we have  $k=4$  and single observations at each failure point, i.e.,  $n_i=1$ ,  $i=1, \dots, 4$ . The number of censored observations in the intervals between failure points are respectively  $m_1=0$ ,  $m_2=2$ ,  $m_3=0$ ,  $m_4=1$ , and  $m_5=1$ . From this, we compute

$$\begin{aligned} r_1 &= 8, & r_1 - n_1 &= 7, & p_1 &= 7/8 \\ r_2 &= 5, & r_2 - n_2 &= 4, & p_2 &= 4/5 \\ r_3 &= 4, & r_3 - n_3 &= 3, & p_3 &= 3/4 \\ r_4 &= 2, & r_4 - n_4 &= 1, & p_4 &= 1/2 \end{aligned}$$

from which we derive

$$\hat{S}(t) = \begin{cases} 1, & 0 \leq t < 0.8 \\ 7/8, & 0.8 \leq t < 3.1 \\ 7/10, & 3.1 \leq t < 5.4 \\ 21/40, & 5.4 \leq t < 9.2 \\ 21/80, & 9.2 \leq t < 12.1 \end{cases}$$

from expression (5). Note how the estimates would differ if we had simply ignored the censored data. In that case, the empirical survival function would give us decrements of 0.25 at each failure time (the empirical survival function is just 1-EDF).

#### LIFE TABLE ANALYSIS

The Life Table method of survival curve estimation is applied in situations when complete information on survival data is not available. The data to which this method applies take the form of interval counts, i.e., only the numbers of individuals who failed or were censored in a given interval are known. Even if complete information is available, however, it is often more convenient to tabulate it in the form of a Life Table.

To illustrate the Life Table method, we use the following example taken from the Connecticut tumor registry. During the years 1946-52, certain information was collected on the Connecticut residents diagnosed as having cancer of the kidney (the data were obtained from Zelen (reference 6)). For each individual, the date of diagnosis was recorded. Each successive year it was noted whether (i) the patient was dead, (ii) the patient was alive, or (iii) the patient was lost to follow-up (LFU) during the 1-year period. The term "lost to follow-up" means that an individual cannot be observed past a certain point in time because he failed to report to the hospital, moved to another city, etc. The analogue of LFU for Navy manpower data might be individuals who go AWOL, for instance.

For those who were diagnosed in 1946, we have the information displayed in table 1 below.

TABLE 1  
PATIENTS WITH CANCER OF THE KIDNEY DIAGNOSED  
IN 1946

Interval (years after diagnosis)	Number alive at start of interval	Number died during interval	Censored observations	
			Number lost to follow- up	Number of withdrawals
0-1	9	4	1	0
1-2	4	0	0	0
2-3	4	0	0	0
3-4	4	0	0	0
4-5	4	0	0	0
5-6	4	0	0	4

In this table, the last two columns represent the censored observations. The term "withdrawal" refers to an individual who is still alive after the period of observation (6 years in this case).

Since the period of observation went until 1952, we have data for 6 intervals for the 1946 cohort group. However, for the cohort group who were diagnosed in 1947 we only have information from 5 intervals, and correspondingly we have 1 less year of observation for each later year of diagnosis. The information for the period 1947-52 is given in table 2.

Suppose we wanted to estimate the probability of surviving 6 years. Since only the 1946 cohort was observed for 6 years, a possible estimate would be  $4/8$ ,  $4/9$ , or  $5/9$  depending on whether or not we discard the one person lost to follow-up, and if not, whether we assume life or death. However, even though the other cohort groups were observed for less than 6 years, these data still contain useful information which we would like to exploit. Tables 1 and 2 can be combined and the data summarized as in table 3.

TABLE 2  
PATIENTS WITH CANCER OF THE KIDNEY DIAGNOSED  
FROM 1947 TO 1951

<u>Year of diagnosis</u>	<u>Interval</u>	<u>Number alive</u>	<u>Number died</u>	<u>Censored observations</u>	
				<u>Number lost to follow-up</u>	<u>Number of withdrawals</u>
1947	0-1	18	7	0	0
	1-2	11	0	0	0
	2-3	11	1	0	0
	3-4	10	2	2	0
	4-5	6	0	0	6
1948	0-1	21	11	0	0
	1-2	10	1	2	0
	2-3	7	0	0	0
	3-4	7	0	0	7
1949	0-1	34	12	0	0
	1-2	22	3	3	0
	2-3	16	1	0	15
1950	0-1	19	5	1	0
	1-2	13	1	1	11
1951	0-1	25	8	2	15

TABLE 3  
LIFE TABLE FOR PATIENTS WITH CANCER OF THE  
KIDNEY

<u>Interval</u>	<u>Number alive at start</u>	<u>Number died during interval</u>	<u>Censored observations</u>	
			<u>Number lost to follow-up</u>	<u>Number of withdrawals</u>
0-1	126	47	4	15
1-2	60	5	6	11
2-3	38	2	0	15
3-4	21	2	2	7
4-5	10	0	0	6
5-6	4	0	0	4

Let  $A_i$  be the event of surviving the  $i$ th interval and  $E_t = A_1 \cap A_2 \cap \dots \cap A_t$  be the event of surviving the years covered by the first  $t$  intervals. Then we may write

$$\begin{aligned} P(E_1) &= P(A_1) \\ P(E_2) &= P(A_2|E_1)P(E_1) \\ P(E_3) &= P(A_3|E_2)P(E_2) \\ &\vdots \\ P(E_t) &= P(A_t|E_{t-1})P(E_{t-1}) \end{aligned} \quad (6)$$

Life Table analysis consists of estimating the conditional probabilities given by (6) and, from these, the probability of surviving  $t$  intervals. The only complications are caused by the censored observations. In order to utilize the censored data, we need to make some distributional assumptions, although these assumptions need only be very weak and do not significantly detract from the nonparametric nature of our estimates.

Before we state these assumptions, however, we need the following definition. The hazard function  $h(t)$  is the conditional probability of a failure in the interval  $(t, t+dt)$ , given survival to time  $t$ , i.e.,

$$h(t)dt = P(t \leq T < t+dt | T \geq t). \quad (7)$$

By use of (7), the survival function can be expressed as

$$S(t) = e^{-\int_0^t h(x)dx}, \quad (8)$$

so that knowledge of the hazard function implies knowledge of the survival function.

We are now ready to state the assumptions behind the Life Table method. These are:

(i) There is a constant hazard  $h_j$  over each interval  $I_j$ ,

(ii) The time of censoring in  $I_j$  for each censored observation is uniformly distributed across the interval.

As long as we take our intervals  $I_j$  to be relatively small, these assumptions are quite reasonable.

If we denote the number of individuals alive at the start of the  $i$ th interval by  $n_i$ , the number who died during this interval by  $d_i$ , and the number censored in the interval by  $m_i$ , then the maximum likelihood estimates of  $p_i = P(A_i | E_{i-1})$  are

$$\hat{p}_i = 1 - \frac{d_i}{n_i}, \quad i=1, \dots, k, \quad (9)$$

where 
$$n_i' = n_i - \frac{m_i}{2}. \quad (10)$$

The numbers given by (10) are called the effective number of observations. In effect, half of the censored observations in the interval  $I_j$  are considered to have survived the interval, while of the other half,  $\hat{p}_i$  proportion of them are assigned as survivors. Censored observations beyond  $I_j$  are, of course, all survivors of that interval. Using (9), the Life Table estimates of  $S(i)=P(E_i)$  may then be obtained as in table 4 below.

TABLE 4

## LIFE TABLE ESTIMATES OF THE PROBABILITY OF SURVIVAL

Interval	Number alive at start of interval	Number died during interval	Number censored	Estimate of $p_i$	$S(i)=P(E_i)$
0-1	$n_1$	$d_1$	$m_1$	$\hat{p}_1 = 1 - \frac{d_1}{n_1}$	$\hat{p}_1$
1-2	$n_2$	$d_2$	$m_2$	$\hat{p}_2 = 1 - \frac{d_2}{n_2}$	$\hat{p}_1 \hat{p}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(k-1)-k$	$n_k$	$d_k$	$m_k$	$\hat{p}_k = 1 - \frac{d_k}{n_k}$	$\hat{p}_1 \hat{p}_2 \dots \hat{p}_k$

If we apply these results to the data of table 3, we compute  $n_1' = 116.5$ ,  $n_2' = 51.5$ ,  $n_3' = 30.5$ ,  $n_4' = 16.5$ ,  $n_5' = 7$ , and  $n_6' = 2$ , from which we obtain the probability estimates given in table 5.

TABLE 5

## LIFE TABLE ESTIMATES FOR THE DATA DISPLAYED IN TABLE 3

Interval	$\hat{p}_i$	$\hat{S}(i)$
0-1	0.5966	0.5966
1-2	0.9030	0.5387
2-3	0.9345	0.5034
3-4	0.8788	0.4423
4-5	1	0.4423
5-6	1	0.4423

What degree of confidence can we place in the Life Table estimates? First, it can be shown that the estimates of  $S(i)$  are unbiased. In addition it can be shown that the variance of  $\hat{S}(i)$  is approximately

$$\text{Var}[\hat{S}(i)] \approx [S(i)]^2 \sum_{j=1}^i \frac{(1-p_j)}{n_j' p_j} , \quad (11)$$

which clearly decreases as the original sample size (and thus all  $n_j'$ ) increases.

Of course, we do not know the true values of  $S(i)$  and  $p_j$  in expression (11), and so we substitute the maximum likelihood estimates  $\hat{S}(i)$  and  $\hat{p}_j$  in their places.

## NONPARAMETRIC ESTIMATION OF A SURVIVAL CURVE WITH COVARIATES

In most Navy manpower problems, we are interested not only in survival probabilities but also in which pre-service or in-service characteristics affect survival. Nonparametric methods for estimating survival curves while adjusting for covariates have been developed only recently, however. At CNA, the most commonly used methods of accounting for covariates have been probit and logit analyses, although these methods are parametric. Since these methods are used so frequently, we shall include a brief discussion of them in this section for the sake of completeness.

A summary of the methods to be discussed in this section is provided below.

- Probit analyses, logit analyses - give point estimates of probability of survival; do not utilize censored data
- Cox regression model - provides continuous estimate of survival curve; utilizes censored observations.

As with methods for estimating survival curves without covariates, this is not an all-inclusive list but does summarize the most commonly used methods.

### PROBIT AND LOGIT ANALYSES

Since both probit and logit analyses have been used extensively at CNA there are a number of CNA publications describing these procedures (see references 1, 2, and 7, for example). For this reason we shall describe these procedures here only briefly.

Suppose a dichotomous response is observed for each individual, such as, the individual stays in or leaves the Navy at a particular point in time. Let this response be denoted by  $Y_i$ , where  $Y_i$  equals 1 if individual  $i$  stays and 0 if he leaves.

Suppose that an individual's pre-service and in-service characteristics are represented by a vector  $Z_i$ .

Then probit analysis models the expectation of  $Y_i$  as

$$E(Y_i) = P(Y_i=1) = p_i = \int_{\beta'Z_i}^{\infty} d\Phi(x) , \quad (12)$$

where  $\beta$  is a vector of coefficients to be determined and  $\Phi$  is the unit normal distribution function. Similarly, the logit model is obtained by substituting  $\beta'Z_i$  for  $\beta'Z_i$

$e^{\beta'Z_i} / (1 + e^{\beta'Z_i})$  for the right-hand side of (12). This closely approximates the normal distribution except at the tails. The likelihood of the observations is

$$L = \prod_{i=1}^n p_i^{Y_i} (1-p_i)^{1-Y_i} , \quad (13)$$

where  $n$  is the number of individuals (assumed to be responding independently of one another). The vector is then estimated by the maximum likelihood solution to (13). The coefficients given by  $\beta$  will tell us how the variables in  $Z$  affect survival. Also, for any set of covariates  $Z$ , substitution of the estimates for  $\beta$  into (12) will give us an estimate for the probability of survival.

Note that the estimate of the probability of survival is for the point in time at which the  $Y_i$  values were calculated. If we wished to approximate a continuous survival curve, we would have to perform a conditional analysis as we did with the Life Table method, estimating a different probit equation at each step. This is very time-consuming, however, and has the additional drawback of giving different estimates for  $\beta$  over time, which may be difficult to interpret.

#### THE COX REGRESSION MODEL

The Cox regression model was first proposed by Cox as a quasi-nonparametric method for estimating a survival curve while adjusting for factors which may affect it. Until recently, it has been applied mainly in the biological and health sciences, particularly in clinical

life studies. The method is termed quasi-nonparametric because only weak assumptions are made about the form of the survival distribution.

Before I proceed to describe the Cox model, recall the definition of the hazard function given by expression (7). The Cox model expresses the hazard function as

$$h_z(t) = h_0(t)e^{\beta'Z} \quad , \quad (14)$$

where  $Z$  is a vector of covariates,  $\beta$  is a vector of unknown coefficients, and  $h_0(t)$  is assumed to be fixed and independent of  $Z$ , but otherwise completely unspecified. Note that  $h_0(t)$  corresponds to the hazard function for the situation when  $Z = 0$ .

Some properties of the Cox model are evident from the formulation given by (14). These are:

- The effects of covariates (i.e., the  $\beta$ 's) are constant over time.
- Differences among the survival distributions of individuals are caused only by changes in  $Z$ .
- Expression (14) is a proportional hazards model; i.e., the hazard for an individual with covariate  $Z_1$  is proportional to that for an individual with covariate  $Z_2$ .

The last property can be seen by writing  $h_1(t) = h_0(t)e^{\beta'Z_1}$  and  $h_2(t) = h_0(t)e^{\beta'Z_2}$ . Then

$$\frac{h_1(t)}{h_2(t)} = e^{\beta'(Z_1 - Z_2)} \quad ,$$

and this expression is independent of time. In terms of the survival functions, this relationship implies (from expression (8)) that

$$S_1(t) = [S_2(t)]e^{\beta'(Z_1 - Z_2)} \quad ,$$

so that one survival curve is a power of the other. In some situations, the proportional hazards property may be too restrictive. The Cox model can be modified,

however, to allow for nonproportional hazards and we shall discuss this modification later.

By keeping  $h_0(t)$  arbitrary, Cox argues that no information about  $\beta$  can be contributed in intervals in which no failures occur, since  $h_0(t)$  might conceivably be zero there. He uses this rationale to justify the use of a conditional likelihood method based on the set of observed failure times  $t_{(1)} < t_{(2)} < \dots < t_{(k)}$  from a sample of size  $n$ . The logic behind using a conditional rather than the usual unconditional method is to obtain a likelihood which is functionally independent of  $h_0(t)$ , which enables us to estimate  $\beta$ .

For each failure time  $t_{(i)}$ , define the risk set  $R_{(i)}$  as the set of individuals who are observed to fail or are censored on or after  $t_{(i)}$ . Then conditional on  $R_{(i)}$ , the probability that the failure at  $t_{(i)}$  is by the individual as observed may be shown to be

$$\frac{h_0(t_{(i)}) e^{\beta' Z_{(i)}}}{h_0(t_{(i)}) \sum_{l \in R_{(i)}} e^{\beta' Z_{(l)}}} = \frac{e^{\beta' Z_{(i)}}}{\sum_{l \in R_{(i)}} e^{\beta' Z_{(l)}}}, \quad (15)$$

where  $Z_{(i)}$  is the covariate corresponding to the individual with observed time  $t_{(i)}$ . Thus the conditional likelihood of the data is formed by taking the product over all failure times of terms such as (15), i.e.

$$L = \prod_{i=1}^k \left\{ \frac{e^{\beta' Z_{(i)}}}{\sum_{l \in R_{(i)}} e^{\beta' Z_{(l)}}} \right\}. \quad (16)$$

Estimates of  $\beta$  can then be obtained by a maximum likelihood solution of (16). This, of course, has to be done by numerical methods.

The model (14) can be extended in a natural way to include cases when the hazard functions cannot be considered proportional to one another with respect to the levels of some factor or combination of factors. Suppose that there are  $s$  strata or blocks (defined by the levels of some factor, such as education). The  $j$ th stratum can be given its own basic hazard  $h_j(t)$  with the dependence on the covariates assumed to be the same for all strata. In the  $j$ th stratum, the hazard function can be written as

$$h_z(t) = h_j(t)e^{\beta'Z}$$

for an individual with covariate value  $Z$ . An analysis of this model gives  $s$  factors of the form (16) to the conditional likelihood of  $\beta$  (one from each stratum).

Now let us consider a method for estimating the underlying hazard function  $h_0(t)$ . The method yields a continuous estimate of the survival function for the case when  $\beta$  is known. Since  $\beta$  is, of course, unknown, it is replaced by its maximum likelihood estimate  $\hat{\beta}$ . We shall take  $h_0(t)$  to be a step function and proceed as was done with the Life Table method, utilizing exact failure and censoring times.

Suppose  $0 < \nu_1 < \nu_2 < \dots < \nu_k$  are fixed, predetermined constants and define the intervals  $I_j$ ,  $j=1, \dots, k+1$ , by  $I_j = [\nu_{j-1}, \nu_j)$ , where  $\nu_0=0$  and  $\nu_{k+1}=\infty$ . We shall assume that

$$h_0(t) = \lambda_j \text{ for } t \in I_j.$$

Next, suppose we observe  $d_j$  failures in the interval  $I_j$  at times  $t_1^{(j)}, \dots, t_{d_j}^{(j)}$  and  $m_j$  censored observations in the same interval at times  $s_1^{(j)}, \dots, s_{m_j}^{(j)}$ .

Let the covariate values corresponding to the failure times be denoted by  $z_1^{(j)}, \dots, z_{d_j}^{(j)}$  and those corresponding to the censored times be denoted by

$z_1^{*(j)}, \dots, z_{m_j}^{*(j)}$ . Define

$$Q_j = \sum_{l=1}^{d_j} e^{\beta' Z_l^{(j)}} (t_l^{(j)} - v_{j-1}) \\ + \sum_{l=1}^{m_j} e^{\beta' Z_l^{*(j)}} (s_l^{(j)} - v_{j-1})$$

$$\text{and } R_j = \sum_{l=1}^{d_j} e^{\beta' Z_l^{(j)}} + \sum_{l=1}^{m_j} e^{\beta' Z_l^{*(j)}} .$$

Then the maximum likelihood estimate of  $\lambda_j$  can be shown to be

$$\hat{\lambda}_j = \frac{d_j}{Q_j + (v_j - v_{j-1}) \sum_{l=j+1} R_l} , j=1, \dots, k \\ \hat{\lambda}_{k+1} = \frac{d_{k+1}}{Q_{k+1}} . \quad (17)$$

From expressions (8) and (14), the maximum likelihood estimate for the survival function is

$$\hat{S}(t) = \exp \left\{ -e^{\beta' Z} \left[ \sum_{i=1}^{j-1} \hat{\lambda}_i (v_i - v_{i-1}) \right. \right. \\ \left. \left. + \hat{\lambda}_j (t - v_{j-1}) \right] \right\} , \quad t \in I_j \quad (18)$$

for an individual with covariate value  $Z$ , where the  $\hat{\lambda}_i$  are given by expression (17).

## COMPARISON OF THE COX AND PROBIT MODELS

In order to examine the efficacy of using the Cox regression model on a future cross-sectional data base, we will compare it with probit analysis on the 1973 recruit cohort of 4-year obligors. A longitudinal data base is necessary to make this comparison, since probit analysis can be applied only to this type of data. Note that we are using the probit estimates as the standard of comparison, not on any theoretical grounds, but on an empirical basis, since probit analysis seems to have given reasonable results when applied in previous CNA studies.

The 1973 cohort will be divided into 2 groups -- one who were promised or received an A-school assignment and another who were assigned to general detail (GENDETs). Separate analyses will be performed for each group, since previous work has shown that the effects of various pre-service and in-service characteristics are quite different with respect to the two groups (see reference 8). When applying the Cox regression model to these data, we will assume that the covariate effects are constant over the 4-year period (the period for which survival curves are obtained). If this assumption is not approximately true, we can expect a poor correspondence between the probit and Cox survival curves. As will be seen, the Cox and probit curves agree rather well, and so the assumption of constant covariate effects seems to be reasonable.

The covariates which we shall consider in our analysis are shown below.

PDEPS = 1 if enlistee has primary dependents  
= 0 otherwise

RACE = 1 if enlistee is non-Caucasian  
= 0 otherwise

MGRP = 1 if enlistee is in mental groups 3 lower or 4  
= 0 otherwise

EDUC = 1 if enlistee has less than 12 yrs of education  
= 0 otherwise

AGE17 = 1 if enlistee is 17 years old  
= 0 otherwise

AGE19 = 1 if enlistee is 19 years old  
= 0 otherwise

AGE20P = 1 if enlistee is 20 or more years old  
= 0 otherwise.

The base group of individuals corresponds to those having a value of 0 for each of the variables listed above.

By performing separate probit analyses at each month considering only those recruits who survived the previous month, we can estimate the effects on survival for each of the 7 covariates in each monthly interval. Since this involves estimating 8 (a constant and seven covariates) x 48 (monthly intervals) = 384 parameters for each group (A-schoolers and GENDETs), we shall not display the parameter estimates here. We show instead the probit survival curves for GENDETs and A-schoolers in figures 1 and 4 respectively. The survival curves in each of these 2 groups are further classified according to quality levels A, B, C, and D, defined below, where MG = mental group and LT12, GE12 denote recruits with less than 12 and greater than or equal to 12 years of education.

	GE12	LT12
MG1-3U	A	B
MG3L-4	C	D

It is not our intention here to analyze the results of these probit analyses in detail, since previous CNA studies have already covered this ground. We would like to note, however, the advantage of obtaining entire survival curves as opposed to point estimates of survival. This is evident on observing the huge attrition rates after only 2 months of service (corresponding to the end of boot camp) in the survival curves for GENDETs. The times at which the losses occur can be of great importance in formulating manpower policies and procedures, whereas merely obtaining a point estimate of survival at 1 year or 4 years, say, would yield very little information on the patterns of attrition over time.

Note that in each of figures 1 through 5, the curves show a sudden dip at 45 months. This occurs as a result of the Navy's early-out program, which allows individuals to leave the Navy or reenlist up to 3 months before their initial term of obligation expires.

The probit estimates of survival were obtained at great computational expense. The computation of the GENDETs' survival curves (corresponding to about 6,000 individuals) took approximately one hour of processing time on a Burroughs B6750 computer. For the roughly 35,000 A-schoolers, however, it took nearly eleven hours of processing time. This amount of computation is clearly undesirable and inefficient. On the other hand, the survival curve estimates from the Cox regression model took only 2 1/2 minutes for GENDETs and 17 minutes for A-schoolers.

Survival curves for GENDETs were calculated with both the proportional hazards model of expression (14) and the nonproportional hazards of the extension model discussed in the previous section. The results are shown in figures 2 and 3. Estimates of the coefficients of the covariates described earlier are shown in table 6 for GENDETs and in table 7 for A-schoolers.

TABLE 6  
COEFFICIENT ESTIMATES IN THE COX REGRESSION  
MODEL FOR GENDETs

<u>Variable</u>	<u>Coefficient</u>	<u>Standard deviation</u>	<u><math>\chi^2</math></u>
PDEPS	0.048	0.053	0.83
RACE	-0.120	0.034	12.46
MGRP	-0.253	0.031	66.59
EDUC	0.207	0.030	47.61
AGE17	0.041	0.034	1.46
AGE19	0.015	0.041	0.14
AGE20P	-0.034	0.043	0.62

TABLE 7  
COEFFICIENT ESTIMATES IN THE COX REGRESSION  
MODEL FOR A-SCHOOLERS

<u>Variable</u>	<u>Coefficient</u>	<u>Standard deviation</u>	<u><math>\chi^2</math></u>
PDEPS	0.058	0.023	6.45
RACE	-0.011	0.021	0.25
MGRP	0.094	0.013	52.56
EDUC	0.197	0.015	172.92
AGE17	0.151	0.015	100.80
AGE19	-0.036	0.015	5.90
AGE20P	-0.077	0.016	23.04

The  $\chi^2$  values given in tables 6 and 7 should be compared with the value 3.84, which is the 0.05 percentile of a  $\chi^2$ -distribution with one degree of freedom. Values greater than 3.84 mean that the coefficients corresponding to these values are significantly different from 0. The magnitudes and directions of the estimated coefficients are similar to those obtained from probit analysis.

Returning to figures 2 and 3, we see that the estimates from the nonproportional hazards version of the Cox model (figure 3) more closely resemble the probit estimates than do those from the proportional hazards model (figure 2). Therefore the assumption of proportional hazards is probably a bit too strong for these data. On the other hand, the assumption that the covariate effects remain constant over time seems to be reasonable. Thus, the nonproportional hazards version of the Cox model appears to estimate survival probabilities rather well, at least for GENDETS.

In light of the results for GENDETS, we decided to use the nonproportional hazards version of the Cox model in estimating the survival curves for A-schoolers. The results are shown in figure 5. Again, the survival curve estimates from the Cox model closely resemble those obtained from the probit model.

From an analysis of the 1973 cohort, we have seen that the Cox model gives reasonable estimates of the survival curves for GENDETS and A-schoolers. Furthermore, the results are more easily interpretable (because of the

assumption of constant covariate effects over time) and computationally more efficient than the probit counterparts. Thus, based on the evidence presented here, the Cox model seems like a potentially useful procedure for estimating survival from cross-sectional data.

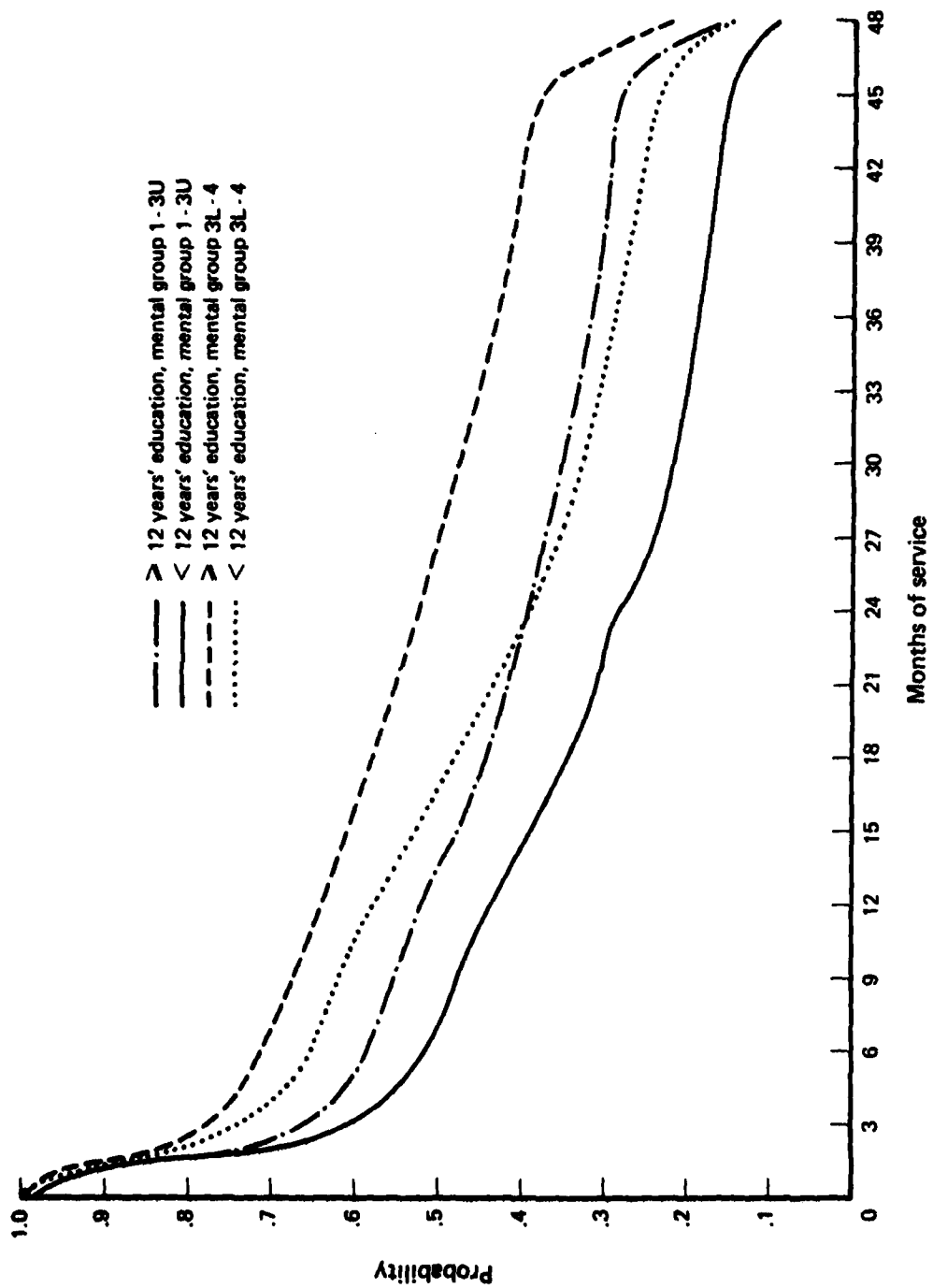


FIG. 1: PROBIT SURVIVAL CURVES FOR 4 YO GENDETS

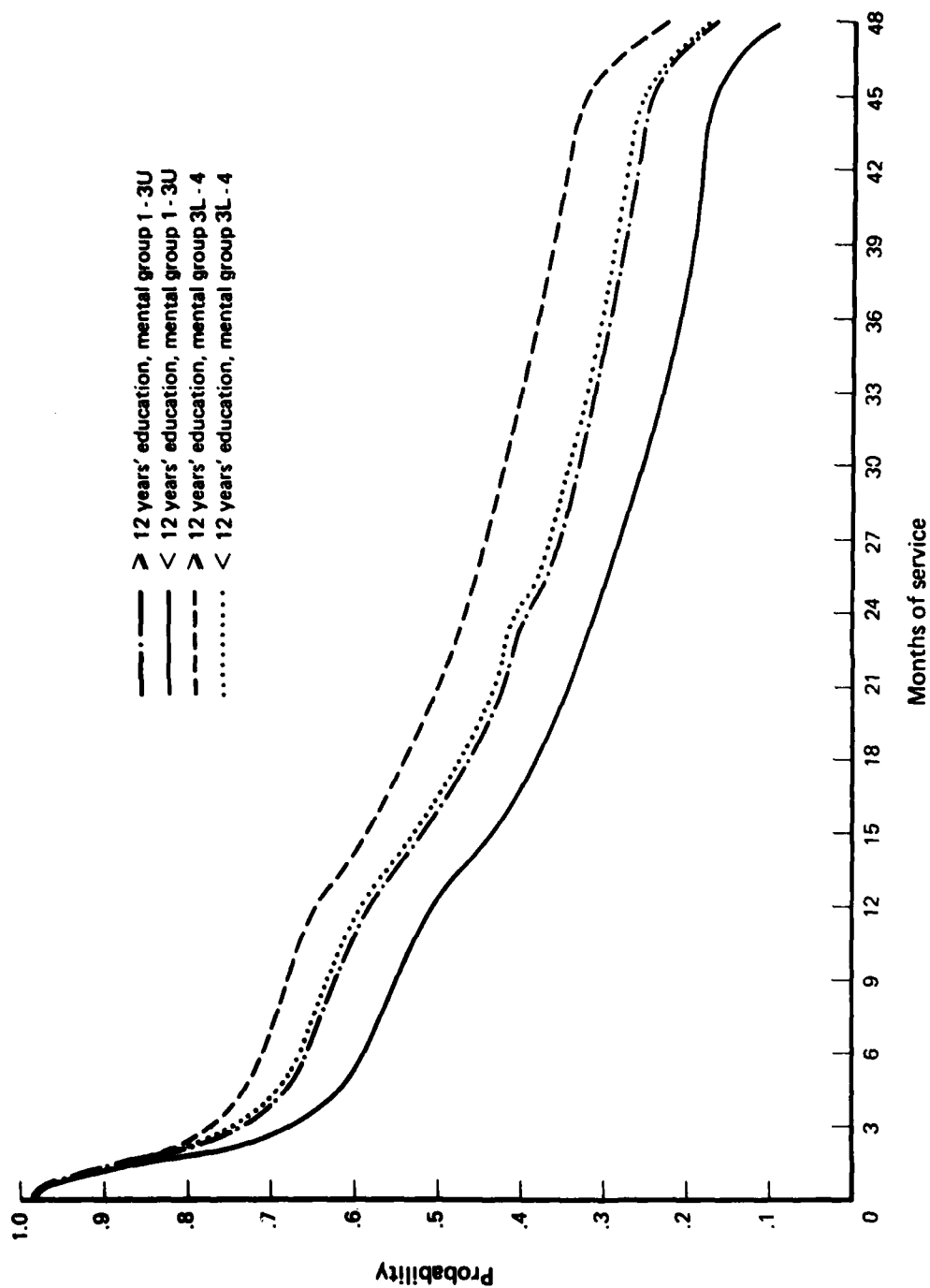


FIG. 2: COX SURVIVAL CURVES FOR 4YO GENDETS-PROPORTIONAL HAZARDS MODEL

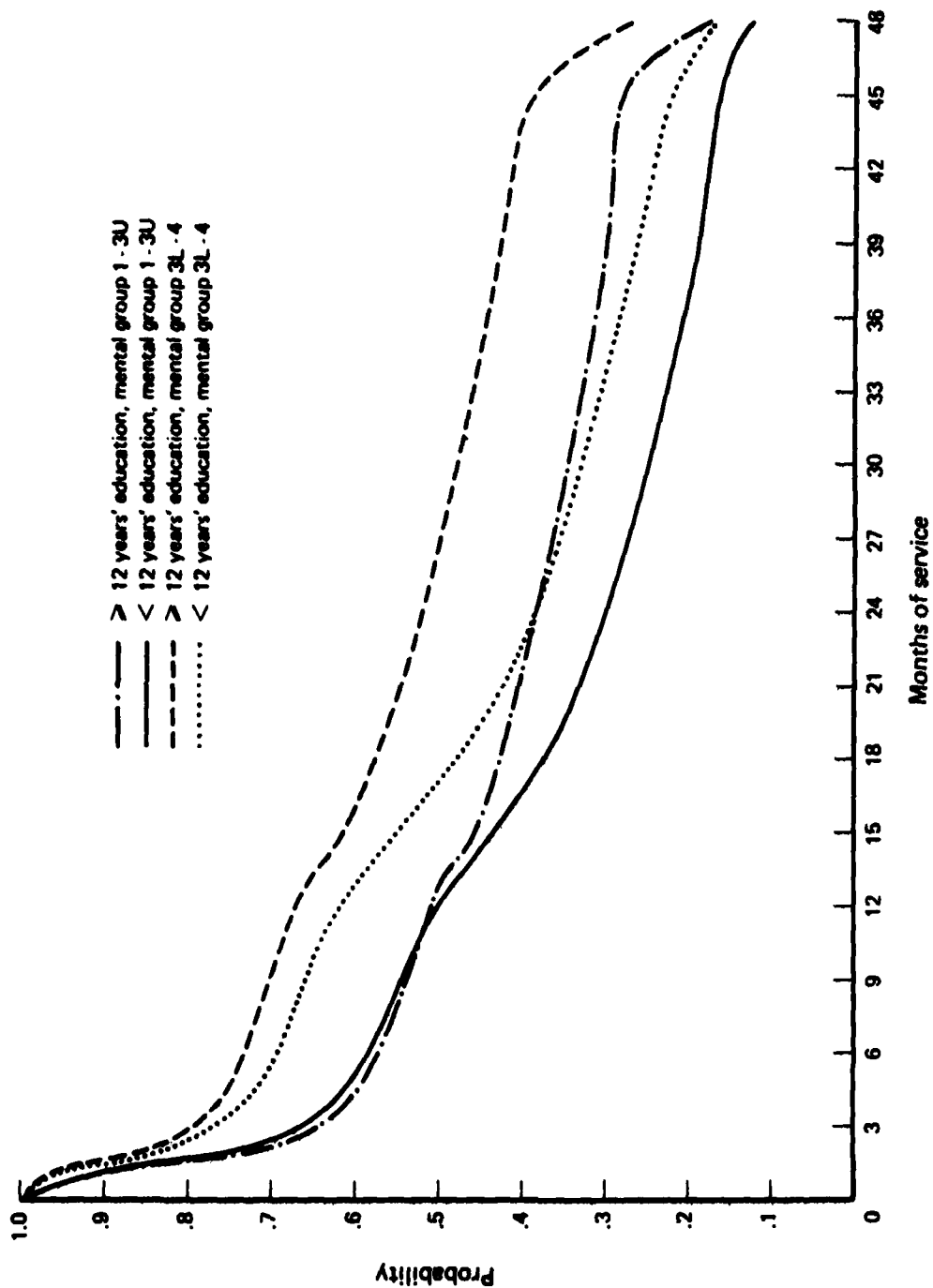


FIG. 3: COX SURVIVAL CURVES FOR 4YO GENDETS-NONPROPORTIONAL HAZARDS MODEL

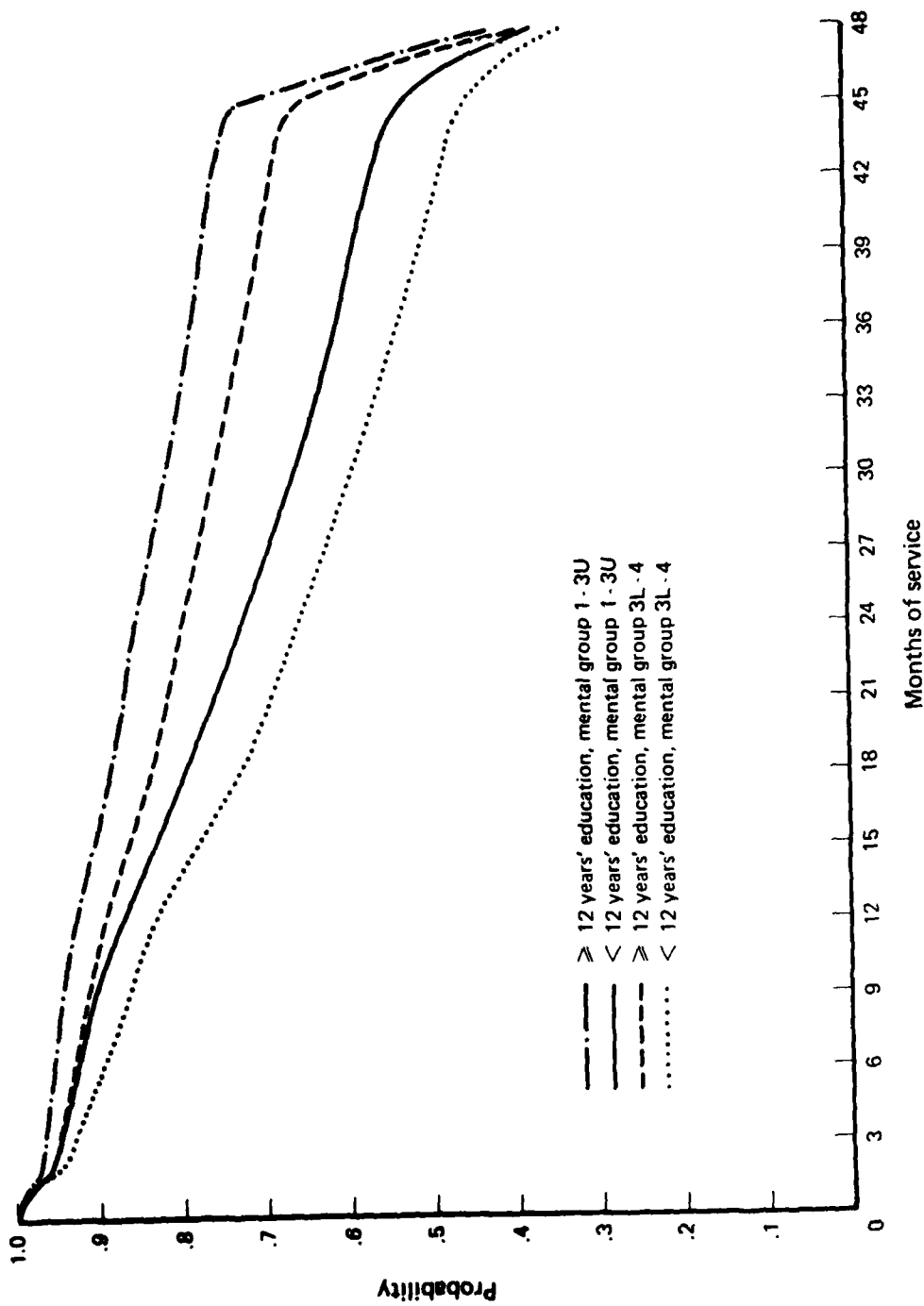


FIG. 4: PROBIT SURVIVAL CURVES FOR 4YO A-SCHOOLERS

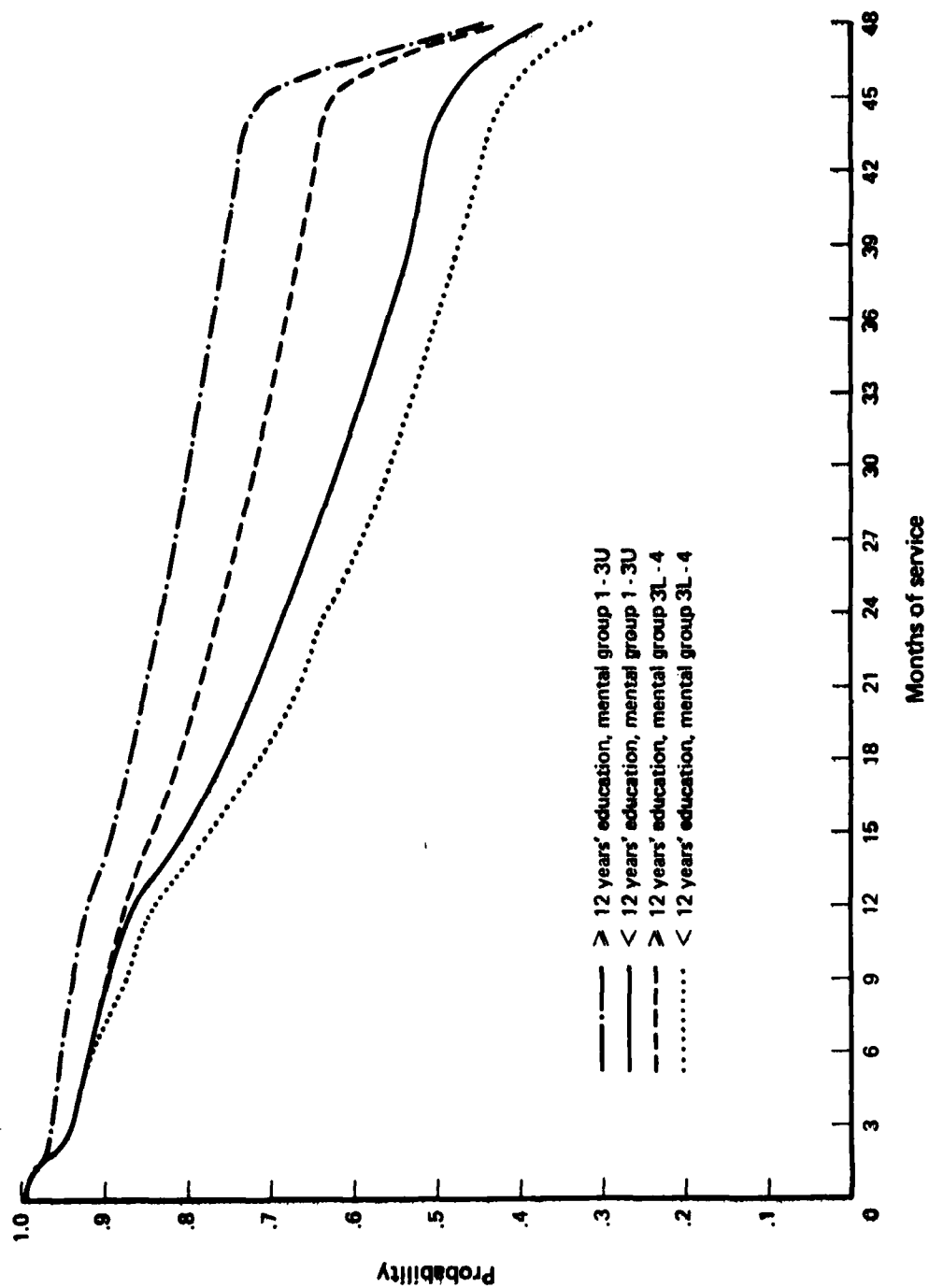


FIG. 5: COX SURVIVAL CURVES FOR 4YO A-SCHOOLERS-NONPROPORTIONAL HAZARDS MODEL

## REFERENCES

1. Center for Naval Analyses, Memorandum (CNA)78-1546, "Predicting Retention of Three-Year Obligor: Application of a Sequential Probit Model," by R.P. Trost, Unclassified, 12 Oct 1978
2. Center for Naval Analyses, Research Contribution 345, "The Prediction of Attrition from Military Service," by J.T. Warner, Unclassified, Sep 1978
3. Cox, D.R., "Regression Models and Life Tables," Journal of the Royal Statistical Society, Series B, Vol. 34, 1972
4. Feller, W., "An Introduction to Probability Theory and its Applications," Vol. 1, John Wiley and Sons, New York, 1968
5. Kaplan, E.L. and P. Meier, "Nonparametric Estimation from Incomplete Observations," Journal of the American Statistical Association, Vol. 53, 1958
6. Zelen, M., "Theory of Biometry," unpublished
7. Center for Naval Analyses, Memorandum (CNA)77-1755, "Predicting Survival of Navy Men from Pre-Service Characteristics After One, Two, and Three Years of Service," by R.F. Lockman, 15 Dec 1977
8. Center for Naval Analyses, Memorandum (CNA)78-1846, "First-Term Success Predictions for Class A School and General Detail Recruits," by R.F. Lockman, 21 Dec 1978